Person Misfit in Attitude Surveys: Influences, Impacts and Implications

David D Curtis
School of Education, Flinders University *david.curtis@flinders.edu.au*

This study of person fit in attitude surveys was undertaken in order to investigate the influence of the inclusion of misfitting persons on item parameter estimates in analyses using the Partial Credit extension of the Rasch measurement model. It was hypothesised that the inclusion of misfitting persons in data sets used for the calibration of attitude survey instruments might compromise the measurement properties of those instruments. Using both actual and simulated data sets, the inclusion of misfitting cases was found to reduce item variance. Several characteristics of both item and person samples were found to influence the proportion of cases identified as misfitting. These characteristics must be considered before removing cases that, according to customary practice, appear to misfit. The residual based misfit indicators that are commonly reported in Rasch analyses, the weighted and unweighted mean squares, appear not to have the generality over all instruments nor the precision required to make clear decisions on the retention or elimination of cases from samples, and there is a need to seek better misfit indicators.

Person Misfit, Attitude Surveys, Rasch

INTRODUCTION

In analysing several attitude survey data sets, up to 25 per cent of cases have been identified as misfitting. In the case of one 40 item instrument that had been developed on a sound theoretical basis, classical item analysis suggested that the scale was coherent with a Cronbach alpha of 0.88, and Rasch analysis showed that all but three items fitted a coherent scale quite well. However, in a subsequent confirmatory factor analysis (CFA), many of the items were shown to have low factor loadings and relatively high error terms, and the fit indices for the hypothesised structure were disappointing (GFI=0.76). These findings led to the search for an explanation of the contradictory outcomes of the analyses. In the review of that instrument, a subset of items was selected, a number of misfitting cases were identified, and their removal resulted in acceptable GFI of 0.93. This observation raised questions about the influence of misfitting cases on the calibration of items and on the integrity of measures derived from the application of the instrument. It also led to a review of other scales that had been analysed previously using the Rasch model. The reviews of these instruments led to the questions: “What are the implications of this high incidence of questionable fit?”; “Does it compromise the inclusion (and exclusion) of items in the scale?”; and “What interpretations can be applied to persons who show either overfit or underfit?” The study has sought to address these and other related questions.

LITERATURE REVIEW

A review of the literature on fit indices revealed that there has been considerable emphasis on item fit, and even in introductory texts, the meanings attached to combinations of fit indices for items are explained well (Bond and Fox, 2001, pp.179-183). However, with the exception of

Wright and Masters (1982) and Wright (1995), the literature has been relatively silent on the implications of fit indices for persons. Wright quoted Rudner et al. (1995) who said:

> Nearly twenty years after Sato introduced his caution index, person fit statistics still seem to be in the realm of potential... The research has been largely unsystematic, the research has been largely atheoretical, the research has not been explored in applied settings. (p.23)

This criticism appears to be harsh, as a considerable body of work has emerged since the late 1980s. However, in most of the studies of fit indices, dichotomous test data have been the main concern. Attitude instruments warrant specific attention for several reasons. They are rarely high stakes activities for respondents and so respondent behaviour may be rather different from that observed in tests, and the number of response categories may interact with misfit indicators.

Two broad sets of issues are canvassed in this review. First, a range of issues identified as being of current interest in educational research and measurement are reviewed. Then literature on indices for the estimation of person fit is examined.

The Centrality of Measurement

Following concerns about the status of measurement in the social sciences, Stevens in 1946, proclaimed that measurement was the "assignment of numerals to objects or events according to a rule." Michell (1997) has shown that such assignment is a necessary but insufficient basis for true measurement as it does not require additivity. Such assignment may produce an ordered sequence, but not an interval one. Wright and Masters (1982, p.3) argued that measurement requires:

- the reduction of experiences to **one dimensional** abstraction;
- **more or less** comparisons among people and items;
- the idea of a **linear magnitude** inherent in positioning objects along a line; and
- a unit determined by a **process** which can be repeated without modification over the range of the variable.

Harwell and Gatti (2001) have argued that the application of item response theory (IRT) is essential to convert the ordered observations that arise from the application of survey instruments to true measures.

In the physical world, measures are objectively observable and conformity with measurement is also observable. For example, if a ruler can be placed adjacent to an object, its length can be measured using that ruler. In the social sciences, constructs of interest are often latent. Instruments, such as questionnaires, are said to make the construct 'observable', but fitness for measurement cannot be directly observed. In the social sciences, fit indices are used to demonstrate compatibility with measurement.

Current Issues in Measurement

Keeves and Masters (1999) identified a number of issues that may impact upon effective measurement using the Rasch model and that are in need of further research. They included dimensionality, significance tests, reliability, and threshold relationships. These issues bear on the study of fit indices, since variations in fit estimates may result from person misfit but also from other characteristics of the items used in the assessment of a trait.

Dimensionality

A vital characteristic of sets of items that purport to measure a construct is that the items are unidimensional. Under Rasch analysis, if all items cohere to form a single scale, unidimensionality may be asserted. However, Kline (1993) claimed that the Rasch model is insensitive to departures from unidimensionality. Often, constructs of interest in the social sciences are complex and are represented by a set of correlated factors. In these cases, it may be difficult to assert unidimensionality. Attitudes are often complex cognitive and conative constructs and theory about them often involves several components. This is certainly the case with the attitude survey instruments used in this study. Whether these constructs constitute several dimensions, or whether the factors operate in concert and comply with the assumption of unidimensionality, must be established. Bejar (1983, p.31) suggested:

> Unidimensionality does not imply the performance on items is due to a single psychological process. In fact, a variety of psychological processes are involved in responding to a set of items. However, as long as they function in unison – that is, performance on each item is affected by the same process in the same form – unidimensionality will hold.

Nonetheless, if a claim is advanced that attitudes are being measured, it is necessary to show that the responses to survey items do comply with the requirement of unidimensionality. Keeves and Masters (1999) proposed that CFA be used to verify item structure and suggested that, in structures with multiple factors, if more variance is explained by the single factor than by the other factors in a nested model, then an assumption of unidimensionality is warranted. They also suggested that both person and item fit should be checked to demonstrate unidimensionality.

Significance

Keeves and Masters (1999) indicated that traditional significance testing, which assumes simple random samples, is inadequate in social science research because simple random samples rarely exist. The problems that ensue when clustered samples are taken to be random, for example underestimation of variance and overestimation of significance, are well known (Kish, 1987). Clustering may apply to items and to the persons who respond to the instruments. While clustering of subjects is well known, the clustering of items is less frequently acknowledged. Keeves and Masters noted that the use of common item stems for multiple responses leads to item clustering and a breach of the assumption of item independence. However, Linacre (1995) has contended that a breach of local independence is less serious than other deviations from measurement model assumptions.

Whatever the cause of departure from ideal model behaviour, for valid measurement it is necessary to quantify this deviation. This is done with a range of fit statistics, and it has been common practice to accept items as fitting if their Infit and Outfit Mean Square values lie within a specified range (often 0.83 to 1.20). Fit statistics are transformed to produce a *t* statistic and the critical values are usually set at ±2. However, Karabatsos (2000) and Smith (1991) have shown that IMS and OMS deviate from normality. Karabatsos also argued that the value of the *t* statistic was sensitive to sample size and that reliance on the *t* statistic could lead to the false detection of misfit. Item fit is often judged on the basis of responses from very large numbers of cases, while person fit is based on a much more modest number of items. Both the magnitude of Infit and Outfit statistics (and other fit indices) and their corresponding *t* values warrant consideration as do the structure of the samples of items and persons that lead to the data sets that are analysed.

Reliability Indices

Keeves and Masters (1999) pointed out that the traditional scale reliability index (Cronbach alpha) is of limited value as an indicator of scale coherence as it is not generalisable. It is dependent upon both the sample of items and the sample of persons used in the calibration. Baker (2001, pp.52-56) provided a very good account of group invariance which leads to the generality of Rasch parameters. Two important instrument parameters are reported in Rasch analyses conducted using Quest (Adams and Khoo, 1999). The Person Estimate Reliability is an indication of the precision of the instrument and shows how well individuals can be distinguished by the instrument and which should be reported routinely in calibrating instruments. Andrich (1982) has shown that this index is virtually identical to the KR-20 or its generalisation, Cronbach alpha. The Item Estimate Reliability shows how well the items that form the scale are discriminated by the sample of respondents. Wright and Masters (1982, pp.90-92) argued that good item separation is a necessary condition for effective measurement.

Threshold Parameters

Attitude survey items usually present multiple response options and so for each item there are multiple item thresholds. Disordered threshold parameters are the subject of disagreement. Keeves and Masters (1999) cited Samejina (1997) who argued that thresholds need not be ordered. Adams and Khoo (1999) identified three types of threshold parameter and suggested that, in order to establish coherence in response patterns, only the Thurstone thresholds need be ordered. In some of the data sets that were analysed in this study, disordered Andrich and Masters thresholds (Taus and Deltas in Quest output) were associated with poor person fit. However, there have also been other complicating factors, such as skewed response patterns and low frequencies of some response options. Keeves and Masters suggested that if thresholds were disordered, it may be possible to redefine categories so that they become ordered. However, Andrich (1995) has advised strongly against this practice, arguing that it violates the assumptions of the Rasch measurement model. This matter is of some importance in attitude surveys, as most use multiple response categories and therefore each item has several threshold levels.

The Measurement of Attitude

In testing, it is common to identify a single construct of interest and to ensure that items are related to that attribute. Multidimensionality is a problem for test developers. For example, where mathematical test items are presented as word problems, the test takers' language comprehension as well as their mathematical skill are involved in the responses that are generated. However, attitudes are often complex cognitive and conative, as well as affective, constructs and theory about them often involves several components. This is certainly the case with the attitude survey instruments used in this study. Whether these constructs constitute several dimensions, or whether the factors operate in concert and comply with the assumption of unidimensionality, must be established.

Weiss and Yoes (1991, pp 72-74) identified four assumptions made of measures. They were:

- if a respondent holds a certain attitude, s/he will respond honestly to an item which taps that attitude;
- the choices that respondents make among response options indicate the strength of the underlying attitude trait that they hold;
- the responses that participants make to particular items are not influenced by the presence of other items in the instrument;
- the pattern of responses to items will conform to a probability function.

These assumptions present challenges to the developers and users of attitude survey instruments as there are many problems that are unique to attitude measurement. There are analogous problems in testing, and guessing, carelessness and prior knowledge of particular items are all acknowledged as threats to test data as a basis for valid achievement measurement (Bond and Fox, 2001, p.178). Anderson (1997, pp.891-3) identified a number of threats to valid attitude measurement, including social desirability, acquiescence, self awareness, irrationality, inadmissibility, self-incrimination and politeness. These influences may have several effects on attitude measures. First, most are likely to decrease overall variation in responses, although some may increase it. Second, they may lead to reduced precision in item and scale parameters. They may also influence the fit of persons to the instrument in that those who hold strong views on some aspects of a construct may provide responses biased by a factor such as social desirability. In addition, and in contrast to achievement tests, responding to attitude survey instruments is rarely a high stakes activity. For this reason, some participants may respond carelessly to an instrument and therefore compromise its calibration. The influences of response behaviours such as these on item and person parameters must be established so that where possible, cases that reveal these behaviours can be identified in data sets and appropriate actions taken by analysts.

Item and Person Fit

Two aspects of item and person fit appear in the literature. One is a technical issue of which fit parameters to use, the distributional properties of those indices, and their sensitivity to departures from expected response behaviours – that is, their ability to detect cases of misfit and to accept fitting cases. The second aspect of item and person fit entails the interpretation of fit indices and the actions that an analyst might take in dealing with cases of misfit.

Technical Aspects of Misfit Detection

The expected response pattern of persons to items is a Guttman pattern, but with added random variance. If a strictly Guttman pattern were to be observed for all respondents, the response model would be deterministic rather than stochastic. The Rasch measurement model is indeed a stochastic one. The ideal Guttman pattern is expected to be approximated but not followed precisely and a range of indices have been developed to quantify the deviation of responses from the Guttman pattern.

Bond and Fox (2001, p.170) reported that Rasch had proposed the use of a chi-square fit statistic to identify how well a data set conformed to the simple logistic measurement model. Once the item difficulty and person location have been estimated from the data set, the expected response for each person to each item can be calculated using an extension of the basic Rasch formula (Wright and Masters, 1982, p.42). The difference between the expected and observed responses is the response residual and is used to compute a chi-square statistic.

Two fit indices are common. The unweighted mean square index, commonly called the Outfit Mean Square (OMS), is the mean of the squared standardised residuals. To calculate the OMS for items, the squared residuals are summed over all persons, and for cases the squared residuals are summed over all items. The information weighted mean square, commonly called the Infit Mean Square (IMS), is the sum of standardised residuals weighted by the variance of the item-person response. For items close to the person location, the variance is at a maximum, while for items that are remote from the person location, the variance is low. Thus, the residuals for well targeted items and persons are weighted up, and for more remote items and persons, the residuals are weighted down. The IMS is therefore less sensitive to unexpected responses for extreme observations and more sensitive to deviations from expectation for well targeted items and persons. These indices are described fully in Wright and Masters (1982, pp.94-105).

Li and Olejnik (1997) compared the performances of five misfit indicators, examining their ability to detect underfit and overfit in unidimensional and two dimensional data sets, and they investigated the influence of test length on misfit detection. They found no correlation between trait estimate and misfit with any of the indicators, which suggests that the concerns expressed by Keeves and Masters (1999) about misfit and trait range (see below) are not a matter of great concern. They also found that the misfit indicators investigated performed equally well under most conditions, although one, L_z was better in detecting underfit in scales that revealed multidimensionality. Li and Olejnik reported that all indicators investigated deviated substantially from a normal distribution. This raises a question about the transformation that is used as a basis for computing *t* statistics for the indicators.

Karabatsos (2000) has been less supportive of the range of misfit indicators that have been used in Rasch analyses. He expressed dissatisfaction with the use of residual based misfit indicators, arguing that misfit indicators based on residuals have unsatisfactory distributional properties. The distributional problem arises from the fact that the residual is the difference between an integer observed score and a non-integer expected score. Karabatsos showed that score residuals are non-linear functions of $(\beta-\delta)$ (cf. Li and Olejnik, (1997) above) and therefore that residual-based fit indices may not indicate the same degree of misfit for different regions of the $(\beta-\delta)$ range. He also argued that the person-item response is used to estimate both β and δ, and as these are used to compute expected values and therefore residuals, it may understate misfit. Karabatsos also showed that the distributions of residuals, IMS and OMS vary with sample size, test length, the person ability distribution, and the item difficulty distribution. With such variations between instruments, and even between applications of the same instrument, different critical values of misfit indicators would be required in order to identify cases of misfit. Further, there are no *a priori* guidelines on what these critical values might be.

Karabatsos also argued that the use of *t* statistics for IMS and OMS is illogical. He cited work by Smith (1991) who showed that the distributions of IMS, OMS, and the corresponding *t* distributions are sensitive to sample size, test length, and person ability and item difficulty distributions. Karabatsos (2000, pp.167-169) went on to show that *t*(IMS) was sensitive to sample size. However, he did this by repeatedly duplicating a data set and producing a set of *t* statistics for each item and each N and showed that the *t* values rapidly diverged. In the original data set, all items had a *t*(IMS) $<|2|$, but with N>8,000 no items were within this range. Unfortunately, the method that Karabatsos used is flawed, because it does not simulate large samples of independent observations drawn from a population. The technique of repeating observations results in no change in the deviation from the mean but with an increase in N leads to reduced error variance and therefore artificially inflated *t* values. A better alternative would have been to identify the ability and difficulty distributions and to simulate data sets of increasing size based on those distributions, and then to look at the *t* distributions. Nonetheless, there is a problem of *t* statistics being sensitive to sample size and test length, and possibly other variables, and this makes the use of the *t* statistic in setting acceptance criteria for items or persons questionable.

Reise (2000) has proposed the use of a multilevel modelling approach in which item responses are thought of as being nested within persons. In this approach, the slope of the person response function is used to indicate response consistency. The person response function slope can also be used as a criterion variable to investigate the effects of between person variables and therefore to investigate the causes of person misfit.

Many others have contributed to the literature on person misfit (Bell, 1982; Linacre, 1998; Meijer, 1996, 1997; Meijer, Muijtjens, and van der Vleuten, 1996; Smith, 1991; Wright, 1995). In particular, Meijer and his colleagues have investigated alternative misfit indicators and the distributional properties of a range of fit statistics. Rudner, Bracey and Skaggs (1996) used person fit indices in an investigation of an achievement test (the National Assessment of Educational

Progress), and found that using person fit indices to identify and remove improbable responses did little to improve an already sound instrument. However, as Karabatsos (2000, p.170) pointed out, most of the work that has been done on person misfit has been directed at the dichotomous responses of achievement tests rather than the more complex situation of attitude survey instruments.

This study was not designed to be a technical evaluation of misfit indicators. Nor was it designed to find alternative indicators of misfit. Rather, it was designed to use readily available fit statistics and to examine some problems that had arisen in practice in the analyses of attitude survey instruments, to explore options for analysts in detecting person misfit when it occurs, and to suggest strategies for managing the analysis of data sets in which person misfit is apparent. However, understanding the technical limitations of misfit indicators is essential in interpreting their operation in practice.

Practical Aspects of Misfit

Keeves and Masters (1999) argued that current practices for the detection of misfit are rather *ad hoc*, and that better procedures are required. They suggested that many misfitting persons are those at the extremities of the trait distributions, and that where $|\beta-\delta|>2$, and they misfit, these cases should be removed from analyses. They also suggested that in analyses, these extreme cases should be weighted down, and those closer to $(\beta-\delta)=0$ should be weighted up to improve estimates. The IMS does place greater weight on the more informative person item interactions that are closer to $(\beta-\delta)=0$.

The issue of person location and fit on a trait measure is an interesting one. If it is argued that an instrument will only have sufficient precision over a limited range ($|\beta-\delta|<2$), then either a range of instrument forms will be required to tap the full extent of the trait range in a population or a form of adaptive measurement, with the attendant concerns of item independence, will have to be pursued. With the common use of polytomous responses in attitude survey instruments, the $|\beta-\delta|<2$ range restriction can be extended to $|\beta-(\delta+\tau)|<2$ since the thresholds show considerably greater spread than the item locations. The influence of person location on fit parameter estimates warrants attention, and at the instrument level, the influence of targeting (or mistargeting) also warrants investigation.

Bond and Fox (2001, pp177-183) suggested ranges of acceptable fit statistics for various test and survey instruments and provide some discussion of the meanings that might be attached to misfit. However, given the concerns raised by Karabatsos (2000) about the distributional properties of residual based fit statistics and about factors that influence them, there is a need to explore their distributions and the sample and item characteristics that might shape them in order to develop advice that is both soundly based and that is useful to practitioners.

METHOD

Two main approaches are followed in this analysis of misfit indicators and of the factors associated with cases being identified as misfitting. First, two data sets that had previously been analysed by the author are reanalysed in order to reveal the numbers of cases that are identified as misfitting. In the discussion below, they are referred to as real data sets. Second, two series of simulation exercises are employed to generate data sets with controlled characteristics.

Analysis of Real Data Sets

The two real data sets selected for the study were the Course Experience Questionnaire (CEQ) that had been administered in 1996 (Johnson, 1997) and the Multidimensional School Anger Inventory (MSAI) (Smith, Furlong, Bates, and Laughlin, 1998) that had been administered in

secondary schools in Adelaide in 2000.[1] Summaries of these data sets are shown in Table 1. A third data set had been selected for the study, but in analyses many items were found to be skewed, some item response categories had quite low frequencies, and many showed threshold reversals. When misfitting cases were removed, item estimates became unstable, and this made the data unsuitable for close simulation (described below). For this reason, no results are reported for it.

Table 1: Summary of real data sets

	CEQ	MSAI36
No. of respondents (N)	51,631	1,400
Number of Items (L)	17	36
Response Categories	5	4
Item Std Dev	0.41	0.76
Person Mean	0.58	-0.25
Person Std Dev	0.94	0.43

The data sets were refined under the partial credit variant (Wright and Masters, 1982) of the Rasch logistic measurement model using Quest (Adams and Khoo, 1999). For the CEQ, eight items were found to misfit and were removed, one at a time, leaving a final set of 17 fitting items. For the MSAI, no items were removed.

In the Rasch analysis, person estimates were generated and exported to an Excel file in which cases were sorted by their misfit parameters. In this way, the numbers of overfitting, fitting, and underfitting cases could be counted easily.

Using the complete data sets, that is without removing cases identified as misfitting, confirmatory factor analyses (CFAs) were undertaken. Subsequently, cases that were identified as misfits were removed and the CFAs repeated in order to provide an indication of data fit to the hypothesised structure with and without the misfitting cases, independent of the Rasch method used to identify case misfits.

Further Rasch analyses were undertaken following the removal of underfitting, overfitting, and finally all misfitting cases. In these analyses, item parameters were tabulated in order to examine the influences on item parameter estimates of the removal of the different types of misfitting cases.

Generation and Analysis of Simulated Data Sets

Two forms of data simulation were used, namely close simulation of real data sets and constrained simulation in which sample and item parameters were controlled.

Close Simulation

In close simulation, the real data sets were analysed and the item parameter estimates (locations and category thresholds) and person estimates were used to generate data sets that were close analogues of the real data. Using the trait level for each person in the original data set, a set of responses to items with parameters identical to those in the real data set were generated using the ConQuest generate command (Wu, Adams, and Wilson, 1998). The data set produced in this way is an analogue of the original real set and conforms to the Rasch model with what may be referred to as an 'expected' amount of noise. The misfit parameters of the closely simulated sample should follow a null distribution, given the item and person sample characteristics, and therefore model

[1] These data were collected by Boman (2002) as a component of his PhD research. Permission to use this data set is gratefully acknowledged.

the proportions of cases that can be expected, under the Rasch model, to overfit, fit and underfit. Such closely simulated data sets provide a useful reference standard for the real data sets with their mixture of fitting cases and those cases that reflect aberrant responses.

Following Rasch analysis, the closely simulated data sets were contaminated with deliberately overfitting and underfitting (extreme) cases and the contaminated data sets were reanalysed. The overfitting cases were generated in Excel using the extended (polytomous) Rasch formula to calculate a set of Guttman patterned responses for a range of abilities from -3.0 to +3.0 logits. (Beyond this range, scores are zero or perfect, carry no information, and parameters cannot be estimated). The overfitting cases follow a deterministic pattern and show minimum noise and maximum signal. The underfitting cases were also produced using Excel and were randomly generated responses to each item. The random responses were noise only and had no signal content, except by chance. The IMS and OMS distributions of the extreme overfitting and underfitting cases were examined. This was done in order to locate optimum critical values for the IMS and OMS statistics to retain as much of the distribution of fitting cases, but to exclude as many misfitting cases as possible.

Constrained Simulation

The real data sets varied in sample size, instrument length, the number of response options, item variance, person variance and the targeting of the instruments. In order to explore the influences of these variables on both scale measurement properties and the proportion of cases identified as misfitting, two constrained simulation series were conducted. In the first, item and person variance and instrument targeting were controlled. Five levels of item variance, from 0.04 to 1.00, were chosen and seven levels of person variance, from 0.04 to 1.96, were chosen. This resulted in 35 combinations of item and person variance. In addition, three levels of instrument targeting were tested by using mean person locations of 0.25, 0.50 and 1.00. In all, there were 105 combinations of the three variables. One data set was generated for each condition. For each data set, the number of cases was set to 500 and the number of items to 20, with five response categories for each item.

In the second constrained simulation series, sample size, instrument length and the number of response options were varied. Samples of size 20, 50, 100, 200, 500, 1000 and 10,000 and with 10, 20, 30, 50 and 70 items were simulated. For each combination of sample size and instrument length, items with three, six, and nine response options were simulated. In all data sets item and sample variance were set at 0.36 and 0.64 and the person mean was set to 0.00. Again, 105 data sets, one for each combination of sample size, instrument length and number of response categories, were generated.

All simulated data files were produced using the ConQuest generate command (Wu et al., 1998).

RESULTS AND DISCUSSION

Because of the large number of data files that were analysed and the many different analyses that were conducted, the detailed results of this study are quite extensive. Only very brief summaries of these results are reproduced in this paper. The simulated data sets, command files and detailed output files are available from the author.

Analyses of Real Data Sets

The Effects on Item Parameter Estimates of Removing Misfitting Cases

Following refinement of the instruments, case parameters including fit statistics, were generated. The proportions of cases that were found to misfit are shown in Table 2. Using criteria modified

after Bond and Fox (2001, p.179), cases were identified as underfitting if the IMS was greater than 1.5 and as overfitting if IMS was less than 0.60. Despite differences in some characteristics of the two instruments, similar proportions of cases were identified as underfitting and overfitting.

Table 2: Summary of cases identified as misfits in real data sets

Data set	Total cases	Underfitting cases	Overfitting cases	Misfitting cases	Fitting cases
CEQ	51,631	5,449	7,093	12,542	39,089
		10.55%	13.74%	24.29%	75.71%
MSAI	1,400	164	181	345	1055
		11.71%	12.93%	24.64%	75.36%

After removing underfitting cases only, overfitting cases only, and then all misfitting cases, item parameters were estimated. Both the dispersion of item locations, that is between item dispersion, and item threshold ranges, within item dispersion, increased following the removal of underfitting cases and decreased following the removal of overfitting cases. The removal of underfits has a stronger influence on parameter estimates, and this is shown when all misfits are removed, as within and between item parameter dispersion increased. These trends are illustrated in Table 3 for item locations and in Table 4 for within item threshold ranges.

Table 3: Summary of the influence of case removal on item locations for real data sets

Data Set	Location	All cases retained	Underfitting cases removed	Overfitting cases removed	Misfitting cases removed
CEQ	Minimum	-0.53	-0.66	-0.50	-0.63
	Maximum	0.64	0.77	0.60	0.70
	Range	1.17	1.43	1.10	1.33
MSAI	Minimum	-1.30	-1.53	-1.23	-1.46
	Maximum	1.13	1.18	1.08	1.12
	Range	2.43	2.71	2.31	2.58

Table 4: Mean Thurstone threshold range under various case deletion conditions for real data sets

Data Set	Threshold Range	All cases	Underfitting cases removed	Overfitting cases removed	Misfitting cases removed
CEQ	$\tau 4$-$\tau 1$	4.05	4.87	3.70	4.46
MSAI	$\tau 3$-$\tau 1$	2.09	2.47	1.92	2.28

The greater dispersion of thresholds within items can be explained with reference to a plot of category probability curves. A sample set of curves is depicted in Figure 1 (produced using RUMM (Sheridan, Andrich, and Luo, 1997) for an item from the MSAI data set). The responses of persons who underfit reveal a higher than expected amount of noise. When they have low trait levels, their expected response would be category '0', but any noise must involve choosing a category greater than this. Thus, the inclusion of underfitting cases must move the threshold up the scale, and the removal of these cases must move it down the scale. The converse holds for misfitting persons with high trait estimates. Thus, the removal of misfitting cases must lead to greater differences between the extreme category thresholds. To a lesser extent this also holds for most intermediate thresholds. For the item shown in Figure 1, which has two intermediate categories, for trait levels where a '1' is the expected response, a noisy response may involve selecting a lower or higher category. But since there are two higher categories and only one lower option, it is likely that the threshold for the expected category will be moved up the scale by the inclusion of misfits and down the scale by their removal. However, such threshold movement should be less than in the case of the extreme categories.

A similar explanation can be invoked for the movements in item locations. When persons respond to items with low locations, most respondents can be expected to select the higher category

options. However, underfitting persons show a greater tendency to select other response categories and as a result the inclusion of their responses leads to a higher estimate of the item location. Their removal leaves the location estimate at a lower position on the scale. The converse argument explains the higher location estimate of the more difficult to endorse items that are located at the positive end of the scale.

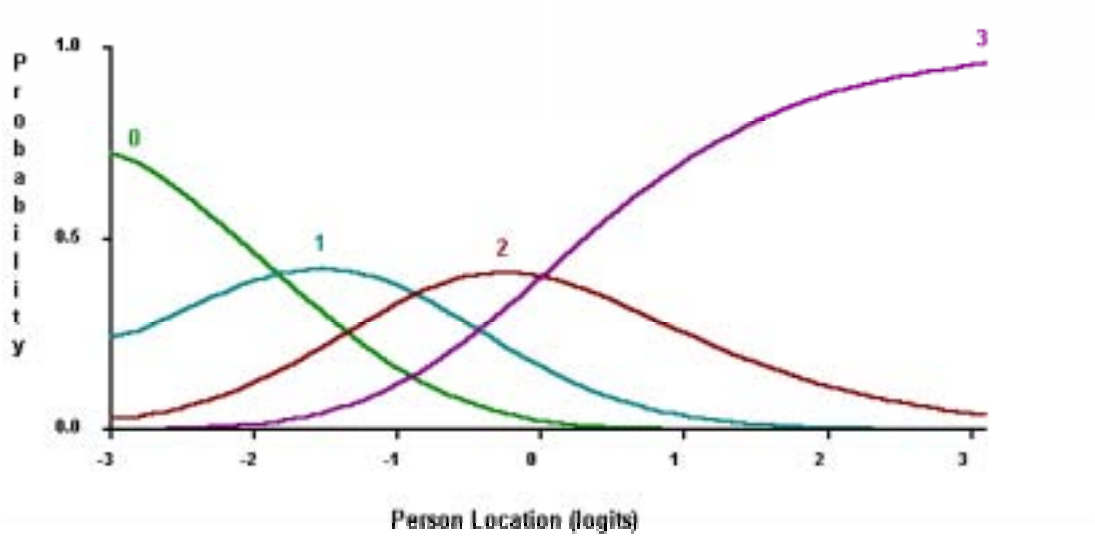

Figure 1: Category probability curves for a four-response polytomous item

Overfitting cases show less than the expected noise. When they are removed, item parameters are estimated using data from the remaining cases which include those that show both an expected amount of variation and those that show greater than expected variation. For the lowest response categories, noise must move the estimated parameters up the scale, and for highest response categories it must shift the estimates down the scale. Thus the removal of overfitting cases results in a compression of item parameter estimates.

The Influence of Misfit on Scale Properties

Wright and Masters (1982, pp.90-91) argued that effective measurement depends upon separation of items along a scale. This evokes the Guttman requirement of having items placed uniformly along a scale and for items to discriminate consistently among respondents. The effect of increasing the dispersion of parameters both within and between items suggests that the removal of misfitting, and especially underfitting, cases should enhance the effectiveness of the scale for measurement. In order to test this, summary scale parameters for data subsets under various deletion conditions were generated and are shown in Table 5.

The item standard deviations show effects consistent with the greater dispersion of between and within item parameters noted above. Item standard deviations increased following the removal of underfitting cases, and decreased after the removal of overfitting cases. In accordance with these changes, person estimate standard deviations increased as item dispersion increased. The item estimate reliability for the CEQ data set was at a maximum value of 1.00. This statistic appears to be influenced by sample size and, with a sample size over 50,000 cases, there is little scope for change. The removal of under- and overfitting cases has little influence on it. Similarly, the case estimate reliability is little changed by the removal of misfitting cases.

Confirmatory Factor Analyses of Real Data Sets

The variances of items and cases carry the statistical information of a data set. In deciding to remove cases identified as misfits, the assumption is made that the level of under- or overfit

apparent in the case, that is the deviation from expectation, makes the case either an outlier or suggests that it does not belong to the population of observations that are consistent with the Rasch measurement model. Setting very rigorous acceptance criteria, such as 0.91<IMS<1.10, is likely to result in the removal of observations whose variances contribute information to the model. Setting very lenient criteria, such as 0.10<IMS<5.00, would lead to the inclusion of random observations with little or no information content. The issue is to decide where to set acceptance criteria. In order to know whether the acceptance criteria used in the analyses of the real data sets were useful, confirmatory factor analyses (CFAs) were carried out on the complete data sets and again on the data sets from which cases, identified in the Rasch analysis as misfitting, had been removed. If the CFA model fit statistics show an improvement it can be assumed that the removal of cases identified as misfitting has resulted in the removal of noise from the data. If the CFA fit indicators are worse, it suggests that cases carrying useful information have been removed.

Table 5: Summary of scale parameters for real data sets under different case deletion conditions

Data Set	Cases	Mean item location (sd)	Item reliability	Mean person estimate (sd)	Case reliability
CEQ	All cases retained	0.00 (0.35)	1.00	0.49 (0.87)	0.88
	Underfits removed	0.00 (0.43)	1.00	0.59 (0.99)	0.89
	Overfits removed	0.00 (0.33)	0.99	0.48 (0.83)	0.87
	All misfits removed	0.00 (0.41)	1.00	0.58 (0.94)	0.89
MSAI	All cases retained	0.01 (0.67)	0.97	-0.21 (0.42)	0.76
	Underfits removed	0.01 (0.87)	0.97	-0.27 (0.45)	0.77
	Overfits removed	0.01 (0.63)	0.97	-0.20 (0.40)	0.76
	All misfits removed	0.00 (0.76)	0.96	-0.25 (0.43)	0.76

A summary of the results of the CFAs is shown in Table 6. The results are mixed. For the CEQ data set, improvements in all fit indices follow the removal of cases identified as misfitting (reduced data sets). However, for the MSAI data set, the reverse is true. The same critical values of IMS were used for both data sets, but for the MSAI set, the CFA suggests that information has been lost while for the CEQ set, it suggests that random variation has been reduced. These results suggest that it may not be sensible to use the same criteria to decide case fit for data sets of difference characteristics.

Table 6: Summary of comparisons of CFA analyses of complete and reduced data sets

Data set	Cases	GFI	PGFI	RMSEA	RMR
CEQ	Complete	0.982	0.629	0.041	0.028
(I factor)	Reduced	0.983	0.630	0.039	0.025
MSAI	Complete	0.937	0.777	0.037	0.032
(4 factor nested)	Reduced	0.931	0.771	0.038	0.032

GFI = Goodness of Fit Index; PGFI = Parsimonious Goodness of Fit Index;
RMSEA = Root Mean Square Error of Approximation; RMR = Root Mean Square Residual

Finding Critical Misfit Values

In order to decide on critical values to separate fitting cases from those that substantially fail to fit the measurement model, it is necessary to explore the distributions of IMS and OMS misfit indicators. The two real data sets were analysed and the case parameters exported to an Excel file

for closer examination. Item parameters were also exported to a file for later use as anchor values. The item parameters and case estimates were used to generate a closely analogous data set that fitted the Rasch model. In addition, two sets of contaminating values were then generated for each real data file. One set conformed to a strict Guttman pattern and the other set were random responses. The former should reveal the IMS and OMS distributions expected of strictly overfitting data and the latter should have distributions representing severely underfitting cases.

The closely simulated data set provides a null distribution of fit indicators, given the characteristics of the instrument and the ability distribution of the respondents, as the data set includes Rasch modelled responses with expected deviations from the ideal deterministic Guttman pattern. The two constructed data sets, the Guttman and random response patterns, provide distributions of fit indicators for worst-case overfitting and underfitting responses respectively. The real data sets include a mixture of fitting and misfitting responses. Critical values of IMS and OMS ascertained in the analysis of the closely simulated and constructed response data sets can then be applied to the real data set in order to find the optimal IMS and OMS values that retain the highest proportion of fitting cases, and therefore retain the maximum amount of statistical information, and exclude the greatest proportions of overfitting and underfitting responses.

The data sets with the contaminated cases were analysed using anchored item parameters derived from the original Rasch analyses. Summary statistics for the distributions of IMS and OMS misfit indicators are shown in Table 7.

Table 7: Summary statistics of distributions of IMS and OMS misfit indicators for real data sets and 'constructed' responses

Data set	Cases	IMS		OMS	
		Mean	Sd	Mean	sd
CEQ	Real	1.02	0.63	1.03	0.65
	Close Sim	1.00	0.33	1.00	0.33
	Guttman	0.23	0.21	0.21	0.15
	Random	2.63	0.64	2.66	0.66
MSAI	Real	1.04	0.47	1.04	0.49
	Close Sim	1.00	0.22	1.00	0.25
	Guttman	0.30	0.20	0.24	0.09
	Random	2.07	0.39	2.32	0.47

Critical Misfit Values for the CEQ Data Set

If the distributions of the misfit statistics were normal, it would be possible to use the means and standard deviations to identify positions on the scale at which optimum separations between overfitting and fitting cases at one end and between fitting and underfitting cases at the other. However, the distributions are not normal. Figure 2 shows the observed IMS distribution for the CEQ data set. It can be assumed that this distribution includes overfitting, fitting and underfitting cases.

In a related study (Curtis, 2003) it has been shown that the distribution of IMS depends upon factors such as item and person variance and the number of response options. Thus, a 'standard' null distribution for IMS under all circumstances cannot be assumed. The distribution of simulated fitting cases (see 'Fits' in Figure 3) may be taken as a null distribution of the statistic. This particular null distribution has been modelled using the person and item parameters estimated in the analysis of the data set and is therefore the best estimate of the null distribution for the conditions under which these data were collected. For this distribution, the points at which 2, 5, and 10 per cent of cases are cut at each end of the distribution are shown in Table 8. The IMS values at any of these cut points could be used to set critical values for accepting responses as fitting. However, it would be much more satisfactory to look also at the characteristics of misfitting cases before deciding on critical values for fit statistics.

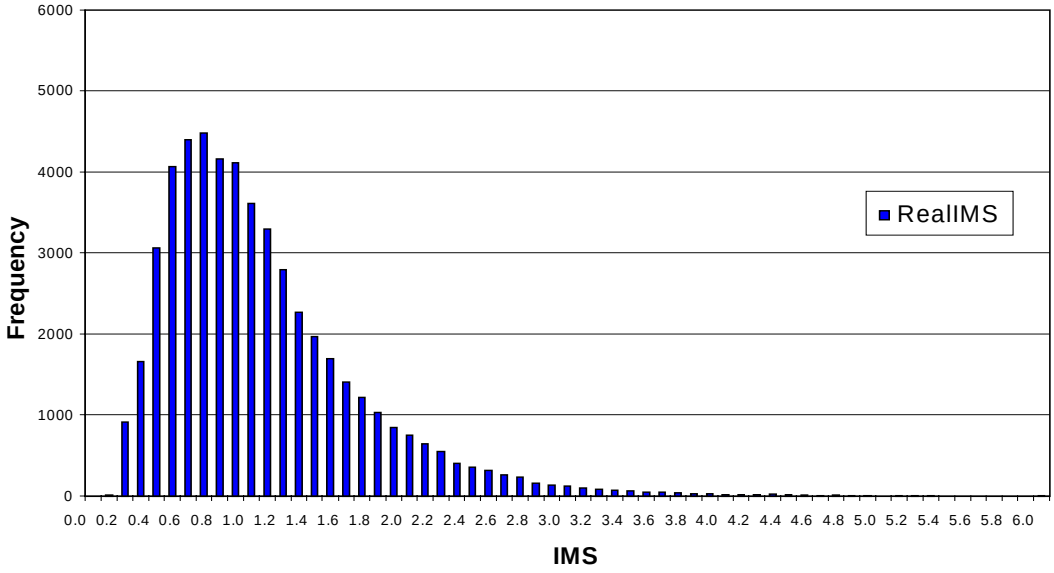

Figure 2: The observed IMS distribution for the CEQ data set

Table 8: Cut points for the IMS null distribution for CEQ data

Distribution	CEQ		MSAI	
tails	Overfits	Underfits	Overfits	Underfits
2%	0.46	1.76	0.56	1.48
5%	0.52	1.58	0.64	1.38
10%	0.59	1.42	0.72	1.28

Figure 3 shows the modelled IMS distributions for simulated overfitting, fitting and underfitting cases for the CEQ data set. In a modelling exercise conducted using Excel, a mix of 7 per cent overfitting, 76 per cent fitting, and 17 per cent underfitting cases most closely approximated the IMS distribution of the real data set. Those proportions of the three categories of cases are reflected in Figure 3.

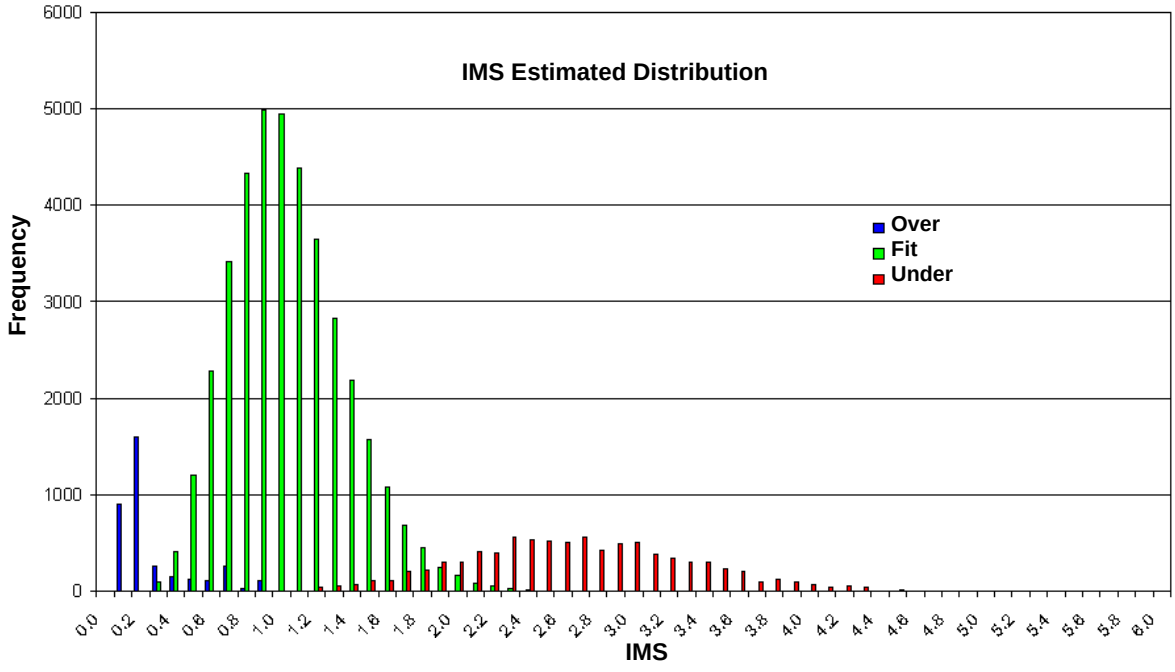

Figure 3: Modelled IMS distributions for overfitting (Guttman), fitting and underfitting (random) responses for simulated CEQ data

The distribution of overfitting cases is bimodal. Inspection of these constructed cases revealed that, while most Guttman responses have IMS values below 0.5, those cases with scores close to zero and perfect have IMS values of between 0.6 and 0.8 and therefore in a range that is generally acceptable. However, most of these cases have OMS values below 0.5. Thus, seriously overfitting cases can be discriminated from closely simulated fitting cases by using both IMS and OMS values. Values of 0.5 for both misfit indicators would remove all but approximately five per cent of Guttman modelled responses, and since overfitting cases make up a modest percentage (seven per cent) in the two data sets investigated in this study, setting critical IMS and OMS values at 0.5 would result in the inclusion of approximately 0.04 per cent of all cases being overfits. This critical value (0.5) would exclude approximately four per cent of modelled fitting CEQ cases and approximately one per cent of fitting MSAI cases. It is worth noting that the use of the two misfit indicators, IMS and OMS, together optimises discrimination between fitting and overfitting cases.

In order to discriminate fitting and underfitting cases, Bond and Fox (2001, p.179) recommended a critical IMS value of 1.4. However, the IMS distribution has been shown to be influenced by several instrument and sample variables, and it may not be useful to propose a single critical value of this misfit indicator. With the CEQ data set, the observed IMS distribution can be modelled best using a combination of 7, 76, and 17 per cent of Guttman, fitting, and random cases. In Figure 3, the distributions of fitting and random cases cross at an IMS value of about 2.0. This may be one convenient point at which separate the two distributions. However, the null IMS distribution for the CEQ data set characteristics revealed that rather lower values of IMS would exclude relatively small proportions of fitting cases, and that a critical IMS value of 1.76 (see Table 8) would exclude only two per cent of fitting cases. Inspection of the distribution of random data revealed that setting this critical value would admit 9.7 per cent of randomly generated cases. But underfitting cases are thought to account for 17 per cent of the observed IMS distribution for the CEQ data, and so this criterion level would lead to a contamination rate of 1.6 per cent. The contamination rates (percentages of random cases that would be admitted) for critical IMS values of 1.58 and 1.42 would be 0.8 per cent and 0.4 per cent respectively.

In seeking to optimise the discrimination of overfitting from fitting cases, it was possible to use both IMS and OMS misfit indicators. However, this is not a useful strategy for underfitting cases as the OMS distributions for fitting and underfitting cases both have greater variance than do the IMS distributions and there is a high correlation between IMS and OMS values above 1.5. Thus, although a clean separation of fitting and underfitting cases is not possible, the use of critical IMS values in the range 1.4 to 1.8 do provide a range of acceptable discriminations. Precisely where to locate the critical value depends upon the proportion of fitting cases that can be eliminated and the proportion of random cases that are accepted. The exclusion of fitting cases represents a potential loss of information, while the inclusion of random cases represents an acceptance of 'noise' that must inflate the standard errors of parameter estimates.

Critical Misfit Values for the MSAI Data Set

The IMS distribution for the MSAI data set is shown in Figure 4. This data set shows less variation than is the case for the CEQ data set (sd=0.47, cf 0.63 for the CEQ data). It is assumed that the observed MSAI data set includes some underfitting and some overfitting cases.

Figure 5 shows the modelled IMS distributions for simulated overfitting, fitting and underfitting cases for the MSAI data set. In a modelling exercise conducted using Excel, a mix of 7 per cent overfitting, 70 per cent fitting, and 23 per cent underfitting cases most closely approximated the IMS distribution of the real MSAI data set. Those proportions of the three categories of cases are reflected in Figure 5.

As was observed for the CEQ data set, the distribution of overfitting cases is bimodal. And, also similarly, most of these generated overfitting cases that have acceptable IMS values, have OMS values below 0.5. Thus, using a combination of critical IMS and OMS indices set at 0.5, it is possible to detect all but approximately 0.15 per cent of these serious cases of overfit. Since overfitting cases make up a only seven per cent of the modelled data, setting critical IMS and OMS values at 0.5 would lead to the inclusion of fewer than one overfitting case in 1,000 and would exclude approximately four in 1,000 modelled fitting cases.

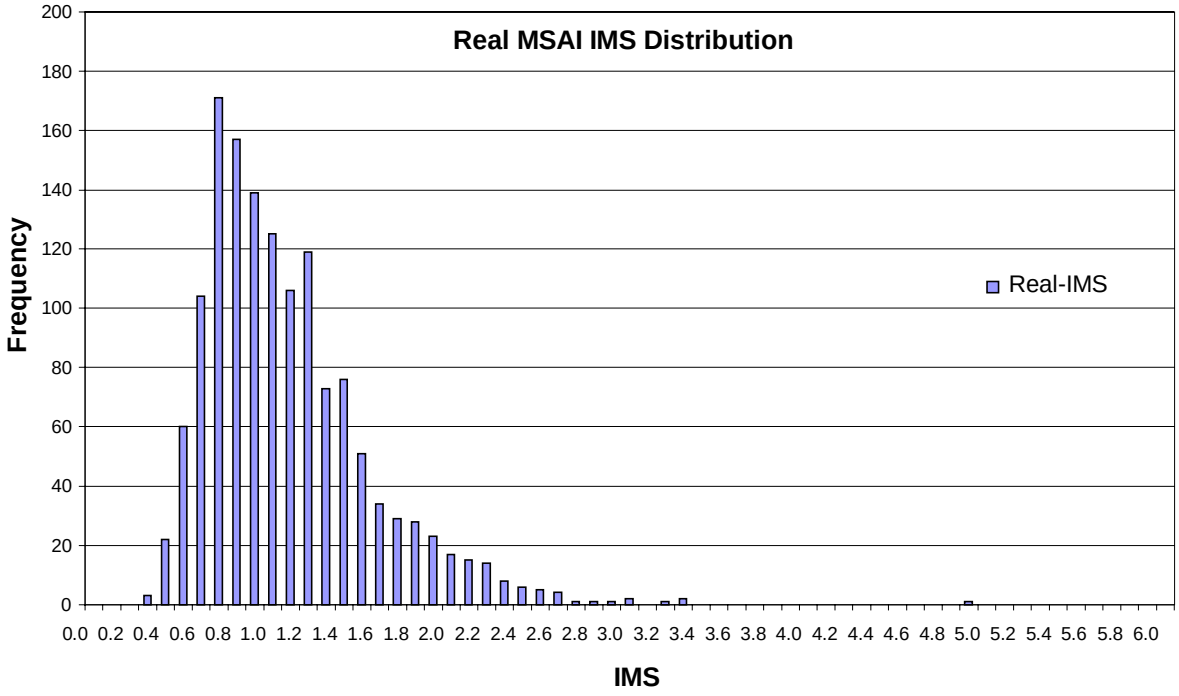

Figure 4: Observed IMS distribution for the MSAI data set

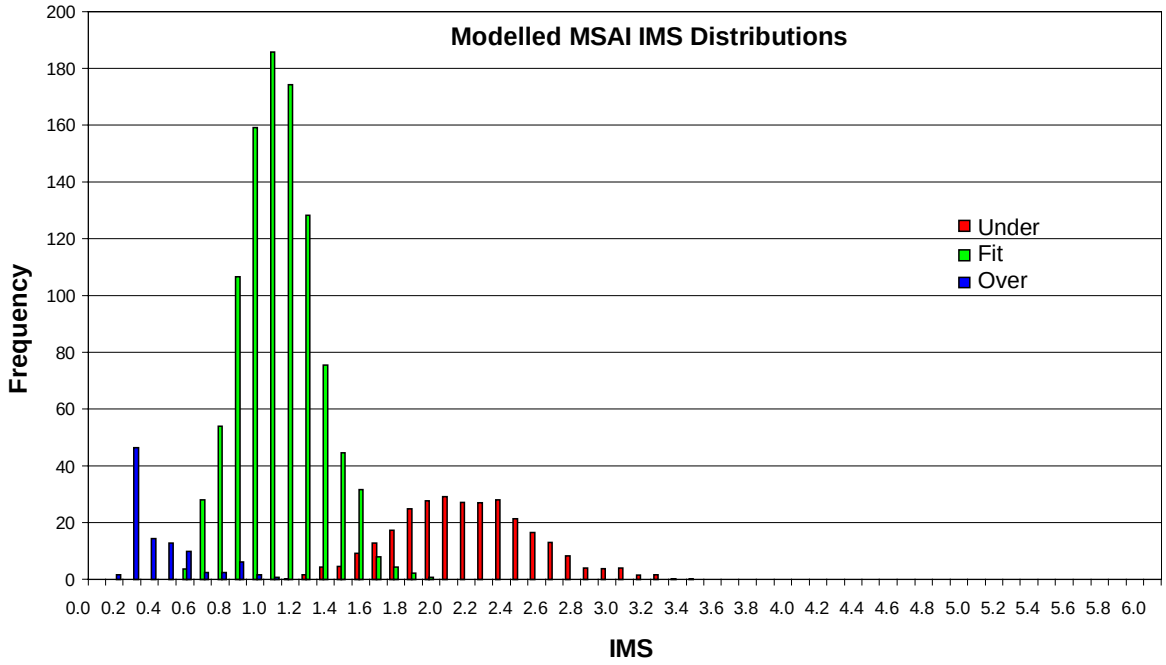

Figure 5: Modelled IMS distributions for overfitting (Guttman), fitting and underfitting (random) responses for simulated MSAI data

The OMS statistic does not assist in the discrimination of fitting and underfitting cases, and the IMS distributions of these two data sources overlap. In Figure 5, the IMS distributions of the

modelled fitting cases and the randomly generated cases cross at approximately 1.6. This appears to be the optimal value at which to discriminate fitting from underfitting cases. At this value, 11.4 per cent of the 27 per cent of randomly generated cases in the sample or 3 per cent of the accepted sample would be contaminated with observations that were mere noise, and 0.7 per cent of the 70 per cent of modelled fitting cases, or 5 per cent of valid cases would be excluded.

Thus critical values for both the IMS and OMS of 0.6 would provide very good discrimination of fitting from overfitting cases. Setting a critical value for the IMS of 1.6, and perhaps even a little higher, would appear to provide the best available discrimination between fitting and underfitting cases for the MSAI data set.

CONCLUSIONS

The research had three purposes. The first was to investigate the influence on item parameters of the inclusion of misfitting cases. The second was to identify the characteristics of samples of both items and respondents that influence the distributions of person fit statistics. A third objective was to attempt to suggest a sound basis for establishing critical values of misfit indicators for attitude survey data.

The analyses undertaken on two real data sets in this study have shown that the inclusion of responses that underfit the Rasch measurement model, and that may reflect carelessness in responding, increase the standard errors of item estimates, reduce the range of item locations on the scale, and reduce the inter-threshold range within items. Thus, the inclusion of misfitting cases compromises the measurement properties of the scale formed by the instrument. Close simulation of the two real data sets, using the item parameters and the distribution of person estimates obtained from the real data sets, has shown that the real data sets appear to include approximately seven per cent overfitting cases and from 17 to 23 per cent of underfitting cases. Together, these findings suggest that it is important to examine person fit as well as item fit in the analysis of data sets and to remove, at least for the purposes of calibration, those cases as well as items, that reveal substantial misfit.

The analyses of constrained simulation data sets has shown that instrument mis-targeting, item and person variance, instrument length, and the number of response options all influence the distributions of the IMS and OMS misfit statistics for persons. This finding suggests that it is possible to provide only broad guidelines about the critical values that might be used in order to discriminate fitting from misfitting cases. From the data sets analysed, it would appear that relatively lenient critical values can be set for the IMS and OMS misfit statistics for persons. The recommendation of an acceptable range from 0.6 to 1.5 (Bond and Fox, 2001, p.179) might be relaxed for attitude survey instruments to 0.5 and 1.6, but that to discriminate well between fitting and overfitting cases, both the IMS and OMS statistics should be used.

In the development of new instruments, it is desirable to establish optimum critical values for these misfit statistics, in order to eliminate as many misfitting cases as possible but to retain the variance of the greatest proportion of fitting cases. The close simulation technique demonstrated in this study provides a method for establishing optimum critical values for misfit statistics. However, these statistics are not able to completely discriminate between fitting and misfitting cases. The effectiveness of alternative misfit statistics that may be developed could be judged on their discriminating power, and the use of real and closely simulated data sets would assist in demonstrating that power.

Before analysts remove cases from data sets, the meanings attached to over- and under-fitting must be understood. Randomly generated responses reveal underfit. In a separate study (Curtis, 2003) pattered responses, the equivalent of so-called 'doodling', also lead to underfit and carelessness may also produce somewhat random responses. Thus underfit, as indicated by high

IMS values, should lead analysts to examine the returns of individuals who evince it. The inclusion of these cases in instrument calibration has been shown to compromise the measurement properties of the instrument. However, in using an established instrument, and especially if anchored item parameters are being used, underfitting cases should be examined closely. It is possible that these individuals are revealing differential item function in the instrument, and this possibility warrants investigation. If the cases are to be excluded from trait estimation, a sound case must be made. This is analogous to the removal of outliers from more common forms of analysis.

Overfit is a clear statistical construct in that it applies to cases that reveal much less variance than is expected in a stochastic response model. However, its practical meaning is less obvious. Random responses that may be attributed to doodling or carelessness and that lead to underfit can be expected. However, for candidates to deliberately set out to overfit the model would require them to apprehend which items were low on the trait and which were high and to generate responses that were overly consistent with this pattern of item locations. Acquiescence and social desirability, identified by Anderson (1997), may also lead to overfit. Another possible source of overfit lies with persons who select intermediate response categories for all items. This might occur if the candidate had decided, without reading the items, to check the central response options. However, it is feasible for candidates to choose these options thoughtfully as the most indicative of their views. Given these conditions, there can be no basis for removing cases that show overfit and that also have a consistent middle-category set of responses. This problem can be ameliorated by ensuring that there are substantial differences in item locations, so that the middle category of one item is located at a trait level that corresponds to either a low or a high response category in other items.

If candidates check all the left-most or all the right-most options, and if there are some reversed scored items, then these returns will reveal underfit. With items that are all worded in the same sense, these response patterns would yield overfit. Thus the inclusion of reverse-scored items in instruments is advantageous for analysts.

The use of information about person fit in Rasch analyses is informative, especially during the instrument development phase. In the analysis of data sets collected using established attitude survey instruments, it is desirable also to examine person fit. Unlike achievement tests, attitude surveys are often anonymous and even if individuals are identified, they are most often low stakes activities for participants. For this reason, researchers might expect that some respondents will be careless and that such data may compromise the effectiveness of the instrument as a measurement tool. Deleting these cases from the analysis may improve the instrument's measurement properties, and removing misfitting cases may lead to more precise scaled data that may then be used as inputs in other forms of analysis.

REFERENCES

Adams, R. J., and Khoo, S. T. (1999). Quest: the interactive test analysis system (PISA Version) [Statistical analysis software]. Melbourne: Australian Council for Educational Research.

Anderson, L. W. (1997). Attitudes, measurement of. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: an international handbook* (pp. 885-895). Oxford: Pergamon.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Educational Research and Perspectives, 9*(1), 95-104.

Andrich, D. (1995). Models for measurement, precision and the non-dichotomization of graded responses. *Psychometrika, 60*(1), 7-26.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). Iowa: ERIC Clearinghouse on Assessment and Evaluation.

Bejar, I. I. (1983). *Achievement testing. Recent advances*. Beverly Hills: Sage Publications.

Bell, R. C. (1982). Person fit and person reliability. *Educational Research and Perspectives, 9*(1), 105-113.

Boman, P. (2002). *Optimism, pessimism, anger, and adjustment in adolescents.* Unpublished PhD Thesis, University of South Australia, Adelaide.

Bond, T. G., and Fox, C. M. (2001). *Applying the Rasch model. Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum and Associates.

Curtis, D. D. (2003). *The influence of person misfit on measurement in attitude surveys.* Unpublished EdD dissertation, Flinders University, Adelaide.

Harwell, M. R., and Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*(1), 105-131.

Johnson, T. (1997). *The 1996 Course Experience Questionnaire: a report prepared for the Graduate Careers Council of Australia*. Parkville: Graduate Careers Council of Australia.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1*(2), 152-176.

Keeves, J. P., and Masters, G. N. (1999). Issues in educational measurement. In G. N. Masters and J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 268-281). Amsterdam: Pergamon.

Kish, L. (1987). *Statistical design for research*. New York: John Wiley and Sons.

Kline, P. (1993). *Rasch scaling and other scales. The handbook of psychological testing.* London: Routledge.

Li, M. N. F., and Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*(3), 215-231.

Linacre, J. M. (1995). Prioritizing misfit indicators. *Rasch Measurement Transactions, 9*(9), 422-423.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2,* 266-283.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*(1), 3-8.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement, 21*(2), 99-113.

Meijer, R. R., Muijtjens, A. M. M., and van der Vleuten, C. P. M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education, 9*(1), 77-89.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88,* 355-383.

Reise, S. P. (2000). Using mutlilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35*(4), 543-568.

Rudner, L. J., Bracey, G., and Skaggs, G. (1996). The use of person-fit statistic with one high-quality achievement test. *Applied Measurement in Education, 9*(1), 91-109.

Sheridan, B., Andrich, D., and Luo, G. (1997). RUMM (Version 2.7Q) [Statistical analysis software]. Perth: RUMM Laboratory.

Smith, D. C., Furlong, M., Bates, M., and Laughlin, J. D. (1998). Development of the Multidimensional School Anger Inventory for males. *Psychology in the Schools, 35*(1), 1-15.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51,* 541-565.

Weiss, D. J., and Yoes, M. E. (1991). Item response theory. In R. K. Hambleton and J. N. Zaal (Eds.), *Advances in educational and psychological testing: theory and applications* (pp. 69-95). Boston: Kluwer Academic Publishers.

Wright, B. D. (1995). Diagnosing person misfit. *Rasch Measurement Transactions, 9*(2), 430-431.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wu, M. L., Adams, R. J., and Wilson, M. R. (1998). ConQuest generalised item response modelling software (Version 1.0) [Statistical analysis software]. Melbourne: Australian Council for Educational Research.